FILOLOGIJA 59, Zagreb 2012

UDK 81'324:811.163.42'374
Pregledni članak
Primljen 5.I.2011.
Prihvaćen za tisak 20.VI.2011.

Vuk-Tadija Barbarić
Amir Kapetanović
Institut za hrvatski jezik i jezikoslovlje
Ulica Republike Austrije 16, HR-10000 Zagreb
vtbarbar@ihjj.hr, akapetan@ihjj.hr

RAČUNALNO OBLIKOVANJE KORPUSA STAROHRVATSKOGA JEZIKA

Rad na projektu *Starohrvatski rječnik* obuhvaća tri glavne faze: tekstološku, računalnu i leksikografsku. Autori se u ovom prilogu bave računalnim oblikovanjem korpusa starohrvatskoga jezika, i to pronalaženjem što učinkovitijega načina kodiranja primarnih podataka i metapodataka za korisnike, ali u okviru realnih mogućnosti projekta.

1. Uvod

Nije minulo mnogo desetljeća otkako su leksikografi prestali ručno ispisivati i abecedirati na tisuće listića kako bi stvorili papirni korpus za izradbu kakva velika rječnika. Računalna tehnologija skratila je i tehnički olakšala mnoge leksikografske pripremne radove koji su se nekoć dugo i mukotrpno obavljali. No, glava leksikografa i danas mora bez pomoći stroja rješavati sve one nedoumice i probleme s kojima su se susretali filolozi doračunalne epohe. Izradba većega povijesnoga rječnika kako je zamišljen *Starohrvatski rječnik* u Institutu za hrvatski jezik i jezikoslovlje, osim donošenja konkretnih odluka na temelju vizije o tome kakav rječnik u konačnici želimo i opisivanja koncepcije leksikografske obrade s popisom vrela i podataka o vrelima, obuhvaća tri glavne faze rada, od kojih su prve dvije pripremne:

1. *tekstološka* (prikupljanje, klasificiranje i kritičko priređivanje starohrvatske jezične građe prema jasno definiranim načelima obradbe)
2. *računalna* (osiguravanje računalne čitljivosti prikupljenih tekstološki verificiranih primarnih podataka stvaranjem digitalnoga arhiva

starohrvatskih tekstova, koji treba prerasti u računalni korpus starohrvatskoga jezika)

3. *leksikografska* (rječnička obradba u jednom od odabranih leksikografskih programa uz potporu stvorenoga računalnoga korpusa).

1.1. Tekstološka faza

Građa za izradbu starohrvatskoga rječnika razdjeljuje se prilikom njezina pripremanja na nekoliko sastavnica. Od 2007. do 2009. godine obrađeni su svi tekstovi prve sastavnice koja obuhvaća starohrvatsko pjesništvo, pa je ta građa pohranjena u digitalnom arhivu i predstavljena 2010. godine u knjižnom obliku.[1] Neversificirani srednjovjekovni starohrvatski tekstovi raspoređeni su u *drugu* (biblijske knjige, lekcionari, psaltiri i molitvenici), *treću* (hrvatska poučna proza), *četvrtu* (hrvatska legendarna proza, priče i romani), *petu* (povijesni i pravni tekstovi) i *šestu* sastavnicu (kratki zapisi i pabirci) i postupno se paralelno pripremaju prema protokolu znanstvenoga istraživanja koje je odobrilo i poduprlo Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske. Time razredba građe nije završena jer se unutar svake sastavnice okupljaju i tekstološki proučavaju starohrvatske verzije i varijante određenih srednjovjekovnih tekstova (a ne, primjerice, čitavi srednjovjekovni zbornici) i klasificiraju se po žanrovsko-tematskom kriteriju. Kako je jedna od glavnih značajki vrela starohrvatske jezične građe tematsko-žanrovska raznovrsnost tropismenih neliturgijskih starohrvatskih vrela tronarječnoga podrijetla (kajkavski samo u natruhama) koji su nastali od 11. do konca 15. stoljeća (pa i onih srednjovjekovnih tekstova što su prepisani do sredine 16. stoljeća), odlučeno je da se svi tekstovi za rječnik transkribiraju hrvatskom latinicom i da se u takvu jezično transparentnom obliku primarni jezični podatci posreduju i u rječniku. Nije prihvatljiv »legalistički« pristup, odnosno nekritičko preuzimanje građe (prijepisi, transliteracije, transkripcije) objavljene u posljednjih 150 godina, neusklađene po kvaliteti i tekstološkim načelima, jer se već prilikom izbora vrela shvatilo da bi tako pripremljeni izvori primarnih podataka bili vrlo loša podloga za leksikografsku obradbu. Građa se nije mogla također pocrpsti izravno iz originalnih rukopisa jer se od samih početaka razmišljalo o rječniku koji bi bio utemeljen na računalnom korpusu tekstova. Stoga kritičko preispitivanje uspoređivanjem dosadašnjih objava različitih inačica tekstova s njihovim izvornicima i obrada tih tekstova latiničnom transkripcijom[2] te transkribiranje do

[1] Kapetanović—Malić—Štrkalj Despot 2010.

[2] Opća transkripcijska načela kao i ona specifična vezana uz srednjovjekovno pjesništvo iznesena su u prvoj knjizi niza *Starohrvatska vrela*. Zašto je odlučeno da se pri-

sada neobjavljenih tekstova (i njihovih inačica) sastavni je dio pripremne tekstološke faze.

1.2. Računalna faza

Građa raspoređena u nekoliko planiranih sastavnica pokazuje obrise strukture korpusa starohrvatskoga jezika koji je u izradi, a tekstovi koji uđu u računalni korpus starohrvatskoga jezika bit će tako označeni da će u konačnici biti moguće u njemu pretraživati podatke u čitavu korpusu ili samo nekim njegovim dijelovima po različitim kriterijima (po pripadnosti određenoj sastavnici, narječnoj podlozi, pismu izvornika, stoljeću itd.). Taj korpus zamišljen je u konačnici kao »računalno čitljiv, označen, jednojezični, nespecijalni statični povijesni korpus pisanoga starohrvatskoga jezika«[3], s tim da odrednica »statični« određuje opći karakter korpusa (stvara se prema unaprijed zadanom i planiranom popisu vrela), ali tim se određenjem ne otklanja mogućnost da neki novi tekst ili nova varijanta već poznatoga (obrađenoga) teksta naknadno u nekoj fazi rada ne uđe u korpus i leksikografsku obradbu. Kako stvaranje tako zamišljena računalnoga korpusa zahtijeva mnogo vremena i posebna stručna IT-znanja i stvaranje/prilagodbu računalnih alata za izvedbu takva korpusa (koji bi po tom pitanju trebao biti usklađen s drugim korpusima u izradi u Institutu za hrvatski jezik i jezikoslovlje), za sada se stvara digitalni arhiv starohrvatskih varijantnih tekstova, koji se u ovoj fazi rada već može pretraživati uz pomoć postojećih konkordancijskih programa. Kritička izdanja koja izlaze u nizu *Starohrvatska vrela* unutar biblioteke *Hrvatska jezična riznica* trenutačan su odraz uređenih sastavnica digitalnoga arhiva i planiranoga računalnoga korpusa. Starohrvatski tekstovi u tim knjigama popraćuju se filološkim osvrtom, kritičkim aparatom i bibiliografskim podatcima, a te će informacije biti iskorištene i kao metapodatci u računalnom korpusu. Kao metapodatak mogu se nuditi i snimke kritičkoga izdanja, pa i originalne snimke pisanih vrela, a snimke bi mogle biti povezane s njihovom transkripcijom.

Stvaranje računalnoga korpusa starohrvatskoga jezika nije preduvjet nego pomoć za leksikografsku obradu. Leksikografska obrada i izgradnja računalnoga korpusa, tehnički gledajući, mogu sukladno teći jer danas postoje takvi leksikografski alati koji omogućuju stvaranje korpusa u svojoj pozadini ili prihvaćanje primarnih podataka prenošenjem iz digitalnoga arhiva ili nekih računalnih alata (npr. iz računalne konkordancije tekstova).

---

mjenjuje latinična transkripcija u pripremi vrela, v. Kapetanović 2007.

[3] Kapetanović 2007:181.

1.3. Leksikografska faza

Leksikografska obrada vrela otpočet će kada kritički priređena građa bude računalno dostupna. Načela leksikografske obrade iznijela je 2002. godine D. Malić, a u kojoj će mjeri biti odstupanja od tako zamišljene obrade, ne samo u tehničkom smislu, još je rano reći. Prije leksikografske obradbe danas je važno izabrati dobar i prilagodljiv računalni program za obradu. U Institutu za hrvatski jezik i jezikoslovlje postoje iskustva s programom *Softlex* (u kojem se izrađuje *Rječnik hrvatskoga kajkavskoga rječnika*), ali se od primjene toga računalnoga programa za izradu *Starohrvatskoga rječnika* odustalo jer ne može ponuditi sve što traži suvremena leksikografska obrada. Stoga je odabran leksikografski program *TshwaneLex*, koji je u uporabi na mnogim uglednim leksikografskim projektima, ne samo europskima. *TshwaneLex* (dalje: *TLex*) jest softver za izradu rječnika koji razvija tvrtka TshwaneDJe HLT[4] još od 2002. Izabran je za izradu Starohrvatskoga rječnika zbog mnogih pogodnosti koje donosi. S obzirom na to da se korpusi u Institutu za hrvatski jezik i jezikoslovlje sastavljaju s pomoću računalnoga jezika za obilježavanje podataka XML[5] u okviru Smjernica TEI-ja[6], veoma je važno da *TLex* podržava *Unicode*[7] — što znači da praktički ne postoje ograničenja u bilježenju slova i dijakritičkih znakova. Također, sam je rječnik (tj. računalna pozadina rječnika) strukturiran u XML-u, što za posljedicu ima iznimnu prilagodljivost potrebama leksikografa.

Utemeljenost korpusa i rječnika na XML-u[8] i unikodu vrijedan je potencijal koji se tek treba istražiti u području povezivanja njihovih baza podataka (u procesu izrade rječnika ili rječnika i korpusa kao gotovih produkata). Međutim, za sada *tlCorpus* — integrirani konkordancijski alat u *TLex-u* — ne može pretraživati korpus u XML formatu, nego samo obič-

[4] *TshwaneDJe Human Language Technology*.

[5] *Extensible Markup Language*.

[6] *Text Encoding Initiative*. Kraticom TEI XML dalje ćemo referirati na takvu uporabu XML-a.

[7] T. Stojanov i T. Portada navode: »Unikod, ključni ISO-v standard suvremenog doba, bavi se digitalizacijom ljudskih znakova (eng. *character*) u pismima (eng. *script*).«, s napomenom da je *Character* »najmanja jedinica pisana jezika koja ima semantičku vrijednost« te da bi taj termin trebalo »jednoznačnije prevesti« (2009:110).

[8] Treba spomenuti da se rad na rječniku ne posprema nužno u XML-datoteku. Tome su razlozi tehničke prirode i uglavnom se tiču brzine dohvata podataka iz datoteke. Ipak, u svakom je trenutku moguće izvesti rječnik u XML. V. Joffe—de Schryver—Prinsloo 2003:244—246. Leksikograf ne mora voditi previše računa o takvim detaljima: »These details are … hidden from the lexicographer, who is thus presented with a consistent and predictable interface for all data storage types« (2003:245).

ne tekstualne datoteke[9] u nekom od unikodnih formata (npr. UTF-8). To nije problematično jer sastavljanje korpusa starohrvatskoga jezika u TEI XML-u ni po čemu nije preduvjet za izradu *Starohrvatskog rječnika*. Dapače, integrirani bi *TLex-ov* konkordancijski alat imao veliku prednost nad takvim korpusom — a to je mogućnost automatskog dodavanja ovjerenih uporabnih primjera u leksikografske članke.[10]

TLex podržava timski rad: nekoliko ljudi s različitih računala može izrađivati rječnik. Štoviše, mogu raditi i na kućnom računalu te bez problema unijeti promjene u glavnu bazu[11], i to bez straha da bi se njihov rad mogao kositi s radom kojega drugoga suradnika — u *TLex-u* je razvijen niz sigurnosnih postavki koje onemogućuju da čiji rad propadne zbog nepažnje. Razvijen je i sustav za dodjeljivanje i uskraćivanje različitih ovlasti u izradi rječnika za pojedine suradnike, što također može biti dobro preventivno sredstvo za očuvanje podataka i strukture rječnika.

Kad se kaže da je *TLex* iznimno prilagodljiv, misli se prije svega na strukturiranje leksikografske natuknice u XML-u. To se čini s pomoću elemenata i atributa kojima se postiže struktura grananja. Elementi u toj strukturi predstavljaju mjesta grananja, a atributi konkretne grane. Prema tome, elementi mogu sadržavati druge elemente i atribute, a atributi samo konkretne vrijednosti. Pogledajmo npr. samo djelić mogućeg leksikografskog članka:

***živsti** živeš nesvrš. …

Svi dijelovi članka koje vidimo pripadaju jednom nadređenom elementu koji možemo nazvati Lema. Vidimo četiri dijela članka. Svaki od njih je atribut koji pripada Lemi: prvi (*) znači pretpostavljenost, tj. nepotvrđenost oblika u korpusu, drugi je lema (ali konkretna, za razliku od Leme), treći kosi gramatički oblik, a četvrti gramatička odrednica. Prilagodljivost se ne odnosi ovdje samo na proizvoljnost i praktičnu neograničenost u dodavanju elemenata i atributa (koji se usto mogu dodatno grafički uređivati da i po tome budu prepoznatljivi u svakom leksikografskom članku) nego i na mogućnost i potrebu pažljiva planiranja strukture članka kako bi se iz što jednostavnije strukture izvuklo što više koristi. Npr. netko bi možda u navedenom primjeru upisao znak * kao dio atributa leme (u konkretnom slučaju to čak ne bi ni poremetilo abecedni poredak u *TLex-*

[9] Dakle, one s ekstenzijom **.txt*.

[10] Ipak, treba primijetiti da su proizvođači *TLexa* osjetljivi na potrebe svojih korisnika te nije isključeno da bi neka iduća inačica mogla podržavati i druge formate, pa i XML.

[11] Bazi se pristupa s pomoću sučelja ODBC (*Object Database Connectivity*) putem SQL-upita.

u). Prikazna bi razina ostala ista, ali leksikograf bi trajno izgubio mogućnost da jednostavno izolira i sagleda sve nepotvrđene primjere ako se za tim pokaže potreba. Osim toga, veoma je važno da se može utvrditi točan redoslijed kojim se prikazuju atributi — to jamči usklađenost svih članaka. Može se čak onemogućiti prikaz pojedinih atributa — tako se iz jedne rječničke baze može dobiti više rječnika za različita ciljana tržišta, čak i s različitim vizualnim rješenjima. Rječnik se može tiskati, ali i objaviti na internetu. Sve to, a i još puno toga više, doprinosi shvaćanju *TLex-a* kao iznimno prilagodljiva alata za sastavljanje rječnika.

Što to znači za *Starohrvatski rječnik*? U principu to da se načela obrade zamišljena i iznesena u Malić (2002) mogu provesti bez većih problema. Ako bude nekih problema u vezi s tim, oni će biti leksikografske, a ne tehničke prirode. Poznato je da je važno dobro osmisliti strukturu članka prije obrade građe. S *TLex-om* se ne ukida ta potreba, premda je on i u tom segmentu prilagodljiv i dopustit će promjenu strukture i u tijeku obrade.

2. Računalno oblikovanje korpusa

U Institutu za hrvatski jezik i jezikoslovlje postoji iskustvo u pripremanju tekstova u formatu TEI XML za Hrvatski jezični korpus (dalje HJK), koje se može iskoristiti i u pripremi starohrvatskih tekstova. Ipak, uočljive su i velike razlike, koje dijelom proizlaze iz odluke da se do primarnih podataka dolazi tekstološkom verifikacijom izvora (za razliku od preuzimanja već izdane građe[12]), a dijelom i iz same prirode izvornika.[13] Nakon tekstološke pripreme mogu se predvidjeti još barem dvije faze:[14]

I. *Uređivanje tekstova u formatu RTF i konverzija u format TEI XML*

Tekst koji je pripremljen za kritičko izdanje ne može u takvu obliku biti transformiran u XML.[15] U ovoj se fazi odvaja kritički aparat od fonološke transkripcije.[16] Većina kritičkog aparata sadržana je u podrubnim bilješkama, pa se to uglavnom odnosi na njih. Bilješke se premještaju na kraj dokumenta. S pomoću regularnih izraza odstranjuju se svi metapodatci koji

[12] U pripremi tekstova za HJK do primarnih se podataka dolazi preuzimanjem elektroničkih izdanja ili, još češće, skeniranjem odabranih vrela i korištenjem softvera za optičko prepoznavanje znakova (tzv. OCR — *Optical Character Recognition*). U idućoj se fazi provjerava rad tog softvera.

[13] V. 2.1.

[14] Na osnovi prvih iskustava bit će izrađen (i poslije nadograđivan) priručnik koji će detaljno opisati svaku fazu.

[15] Može, ali to bi bitno usporilo put od RTF-a do željenog valjano strukturiranog XML-a.

[16] V. 2.3.

su sadržani u samom tijelu teksta, a ne žele se kodirati (npr. obrojčenje svakog petog stiha u pjesmama), ili se mijenja njihova pozicija (npr. obrojčenje folija u rukopisu/spomeniku), ili ih se pak mijenja u unaprijed određen i propisan jednoznačan niz znakova koji će se mijenjati u idućim fazama (npr. oznake za rekonstruirani tekst). Promjene se provjeravaju usporedbom s kritičkim izdanjem.

Na koncu se provodi konverzija RTF-a u format TEI XML u programu za obradbu teksta *Writer* (paket *OpenOffice*) s pomoću XSLT skripte.[17]

II. *Uređivanje tekstova u formatu XML* (*u aplikaciji* ***Oxygen XML Editor***[18])

U toj se fazi mijenjaju nizovi znakova (spomenuti u I. fazi) u elemente i atribute XML-a. Dokument se strukturno uređuje te se uređuju metapodatci u zaglavlju dokumenta. Također se posebnim atributima može povezati tekst s preslikom izvornika. Eventualno se otklanjaju naknadno uočene pogreške ili one neočekivano nastale u konverziji u format XML. Završnu provjeru sintaktičke strukturiranosti dokumenta i ovjerenosti prema specifikaciji TEI-ja radi sam softver.

Na koncu se tekst mrežno objavljuje uz pomoć poslužiteljske aplikacije *Philologic*[19].

2.1. Očekivani problemi i moguća rješenja

Priroda izvornika može dovesti do situacije u kojoj se mora improvizirati kako bi korisnik na prikaznoj razini došao do očekivanih podataka. Naime, *Philologic* sigurno prepoznaje samo elemente i atribute TEI Litea[20], a u njemu nisu sadržana inače prilično detaljno razrađena poglavlja TEI-ja presudna za kvalitetno označavanje rukopisa (*10 Manuscript Description*, *11 Representation of Primary Sources*, *12 Critical Apparatus*)[21]. Razvojni će se tim *Philologica* teško u bližoj budućnosti odlučiti na potrebne nadogradnje, pa jedino ostaje prilagoditi se (što ne isključuje i razvoj vlastitih prilagodba *Philologica*) ili razmotriti mogućnost odabira koje druge poslužiteljske aplikacije.

[17] Navedeni je softver pogodan jer je besplatan, a XSLT skripta dostupna je na stranicama *http://www.tei-c.org/Tools/Stylesheets/* i može se slobodno prilagođavati, premda to traži naprednija informatička znanja.

[18] Ta je aplikacija komercijalna i unatoč tome veoma raširena među korisnicima TEI-ja.

[19] V. *http://sites.google.com/site/philologic3/home.*

[20] TEI Lite je pojednostavnjena verzija TEI-ja. Sadržava elemente i atribute za koje se vjeruje da mogu ispuniti i do 90 % potreba za kodiranjem. Vidi: *http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html*.

[21] V. TEI 2007:279—376. Osim toga, sve je veći interes za iscrpno kodiranje rukopisa. Vidi npr. stranice posebnog tijela TEI-ja *Manuscript Special Interest Group* na *http://www.tei-c.org/SIG/Manuscripts/*. Oni predlažu vrlo razrađen sustav kodiranja.

Pod spomenutom improvizacijom podrazumijevamo ponešto promijenjene uporabe elemenata i atributa TEI-ja, tj. takve koje bi odstupale od preporučenih (ili još češće takve koje imaju puno bolje i preciznije ekvivalente u prije navedenim poglavljima Smjernica TEI-ja)[22], a ponajviše u svrhu kodiranja metapodataka o rukopisima. Za sve ono što bi inače trebalo biti definirano npr. elementom <msDesc>[23] u okviru elementa <sourceDesc>[24], trebalo bi naći alternativna rješenja, i to na taj način da korisnik korpusa ne bude svjestan problema. Što se tiče strukturiranja[25] podataka u bilješkama, tj. u kritičkom aparatu, malo što se može učiniti bez uporabe TEI-jeve dobro razvijene sheme za kritički aparat. Međutim, čak i da je dostupna, teško da bi to dovelo do odluke o njezinoj široj primjeni. To bi zahtijevalo previše vremena, a nije sasvim jasno koliko bi bilo korisno.[26]

2.2. Prioriteti i posljedice

Zbog svega navedenog prvi na popisu prioriteta bit će jednorazinski[27] prikaz fonološke transkripcije[28], a sve više od toga smatrat će se dodatnim ostvarenjem. Kako bi se transkripcije starohrvatskih tekstova mogle provjeravati, pružit će se ne samo iscrpni podatci o izvorniku koji mogu korisniku pomoći u pronalasku izvornika (institucija u kojoj se čuva, signatura i sl.) te lokacije teksta u izvorniku (folijacija) nego će se omogućiti i uvid u preslik izvornika (za one rukopise za koje to bude bilo moguće). Također će biti korisna uputa na bibliografske jedinice koje sadržavaju prijašnja čitanja izvornika (povijest čitanja teksta). Tako korisnik jednostavno može doći do svih potrebnih podataka za samostalnu interpretaciju. To je drugo

[22] Naposljetku bi to moglo značiti nestrukturirane podatke na onim mjestima gdje je inače preporučljiva strukturiranost. V. bilj. 25.

[23] Element <msDesc> (manuscript description) sadržava različite podatke o rukopisu.

[24] Element <sourceDesc> (source description) sadržava različite podatke o izvoru teksta, ali tek uporabom elementa <msDesc> mogu se kodirati podatci specifični za rukopise.

[25] Strukturirani su podatci oni koji su kodirani, tj. koji su eksplicirani za računalo. Moglo bi se reći da su strukturirani podatci oni koji su "jasni" računalima.

[26] Prije svega, očekuje se da će se kritičkim aparatom služiti samo rijetki, a ne očekuje se da sam kritički aparat postane predmetom istraživanja.

[27] Kao uzor u organiziranju kodiranja rukopisa svakako nam može služiti projekt MENOTA (Medieval Nordic Text Archive, na: *http://www.menota.org*) sa svojim dobro razrađenim smjernicama. MENOTA preporučuje prijepis teksta do tri razine: »faksimilska razina«, »diplomatička razina« i »razina normalizacije«. Posljednja se može usporediti s našom fonološkom transkripcijom.

[28] To je i logično jer je tekstološki rad temelj svim drugim planiranim poslovima te je apsolutni prioritet u okviru cijeloga projekta.

na popisu prioriteta, a tek je na trećem mjestu omogućavanje uvida u kritički aparat bez strukturiranja ili s minimalnim strukturiranjem podataka.

Zbog ostvarivanja predloženih prioriteta bit će manje vremena za kodiranje neprioritetnih podataka. U metapodatcima se neće poklanjati pažnja fizičkim karakteristikama rukopisa (visina stranica, širina i sl.). Korpus se neće, barem ne prije izrade rječnika, morfosintaktički označivati. Lematizacija je vrijedna razmatranja, ali tek će trebati ocijeniti koliko bi od nje bilo koristi za izradu rječnika.[29]

2.3. Struktura korpusa[30]

Veoma je važno shvatiti da korpus neće biti prikaz tiskanih kritičkih izdanja. Ona su samo put njegove geneze — jednom kad nastane bit će neovisan o njima, ali zato i dalje ovisan o vjernosti jeziku izvornika.[31] Dakle, očekuje se da se korpus može u detaljima naknadno mijenjati, pogotovo ako se odluči odstupiti od nekih načela.

Uvodne studije kritičkih izdanja neće biti dio korpusa. One će biti vrelo za metapodatke, ali nikako ne smiju činiti tijelo[32] teksta. Osvrte o pojedinim tekstovima i uvodne studije možemo uvrstiti kao vrela u korpus suvremenoga standardnoga jezika. U tom je smislu problematično i kodiranje kritičkih bilježaka. Dosadašnji HJK nije imao problema s prezentacijom podrubnih bilježaka jer je očekivano stanje da se i u glavnom tekstu i u bilješkama pronalazi tekst koji može biti predmetom istog istraživanja — tj. jezik na istoj sinkronijskoj razini.[33] Stoga može sebi dopustiti da ne obraća tome pozornost, ali starohrvatski korpus pak ne može jer kritičke bilješke ne pripadaju starohrvatskom jeziku. Stoga bi možda najsigurnije

[29] O prednostima i ograničenjima neoznačenih dijakronijskih korpusa v. Schulte 2009. Schulte analizira mogućnosti pretrage takvih korpusa s pomoću regularnih izraza, i to na više razina (leksička, morfološka, sintaktička te u takvu formalnom pristupu nedohvatljiva semantička razina). Logično, rezultati su pretraga najbolji na prvim dvjema razinama, što nam i odgovara. I K. Kučera ističe: »Vocabulary is arguably the area where the contribution of diachronic corpora to the mapping of time continuum of a language is most obvious.« (2007:6).

[30] Ovdje se ne misli na uzorkovanje koje se radi da bi se postigla reprezentativnost korpusa. Starohrvatski korpus smatrat će se, sa svim posljedicama koje to nosi, u potpunosti reprezentativnim za starohrvatski jezik u tom smislu da će sadržavati sve tekstove napisane tim jezikom (ili je to bar krajnji cilj). O problemima reprezentativnosti dijakronijskih korpusa v. npr. Kučera 2007:1—2 ili Gau 2005:20—25.

[31] Posljedica je davanje izrazite prednosti logičkoj strukturi izvornika (koja je nužno posljedica interpretacije) — od fizičke strukture bit će vjerojatno kodirana samo folijacija.

[32] Tj. ne smiju se nalaziti u granicama elemenata <body> i <text>.

[33] Moguće je da to nije u svakom slučaju tako, ali to je statistički zanemarivo.

rješenje bilo "fizičko" izmještanje kritičkih bilježaka u odvojenu XML datoteku. Pri tome se naravno mora paziti da one ostanu povezane s točno određenim mjestima u korpusu.[34] Međutim, odabir je ipak uvjetovan ovisnošću o preporučenim rješenjima *Philologica*. U primjeni na naš korpus to bi izgledalo otprilike ovako:

```
<v n="9">I vi, plačne udovice<ref type="note" id="ref9"
target="n9" n="9"/></v>
<v n="10">privelike žalostnice</v>
<v n="11">ke ste muže pogubile,</v>
…
```

Na drugom mjestu u dokumentu:

```
<div1 type="notes">
<head>Bilješke</head>
<note id="n9" place="foot" target="ref9">Graf. <hi
rend="italic">vdouice</hi>.
…
Može se čitati i <hi rend="italic">vdovice</hi>. Vode-
ći se brojem slogova u stihu, u transkripciji pretpostavlja-
mo vokalizaciju prijedloga/prefiksa <hi rend="italic">vъ</
hi>(<hi rend="italic">-</hi>)u <hi rend="italic">u</hi>(<hi
rend="italic">-</hi>).</p>
</note>
```

Najvjerojatnije će svaki pojedini tekst biti kodiran u vlastitoj datoteci, iako ne bi bilo loše rješenje da se sve varijante pojedinog teksta kodiraju u istoj datoteci — o tome svakako treba razmisliti. Sve datoteke bit će povezane metapodatcima s pomoću kojih će ih se u svakom trenutku moći precizno identificirati i, što je najvažnije, sortirati prema različitim kriterijima. To će korisniku omogućiti da pretražuje samo željene tekstove. Osim XML datoteka, nadamo se da će postojati i baza slika izvornika.

3. Zaključak

Na ovom mjestu možemo primijetiti da starohrvatski korpus dvije bitne tendencije povezuju s dijakronijskim korpusom ruskog jezika u Regensburgu:[35] težnja da se pruže lingvistički relevantni podatci i fleksibilnost u kodiranju podataka. Fleksibilnost je osobito važna jer konstantno omogućuje prilagodbu praktičnim potrebama u kodiranju, a prilagodbe će sigurno rasti s brojem tekstova. Poveznica s dijakronijskim dijelom

[34] Metoda o kojoj je riječ opisana je u Smjernicama TEI-ja u poglavlju *12.2 Linking the Apparatus to the Text*. Riječ je o metodi koju nazivaju »location-referenced«. Ona se može koristiti unutar iste XML datoteke ili kao veza između različitih datoteka.

[35] V. Meyer 2005.

Češkog nacionalnog korpusa jest transkripcijski princip i težnja punoj reprezentativnosti.[36] Za sada se čini da starohrvatski korpus donekle izdvaja inzistiranje na koncepciji pripreme tekstova na osnovi izvornika (tj. njihovih preslika), ali o mjestu korpusa starohrvatskog jezika među sličnim projektima u slavenskom svijetu moći će se govoriti tek kada se pojave prve njegove inačice. Jedno je sigurno – (staro)hrvatski se jezik mora izboriti za svoje mjesto među njima.

Suradnici na projektu *Starohrvatski rječnik* jesu filolozi, njihov je broj ograničen, a korpusna lingvistika relativno novo znanstveno područje koje se intenzivno mijenja. S obzirom na prioritet tekstološkoga i leksikografskoga rada može se očekivati da ćemo na računalni korpus čekati još neko vrijeme, ali će se to vrijeme iskoristiti za pažljivu i dosljednu njegovu konstrukciju. Nešto je realnije doskora očekivati objavljivanje pojedine njegove sastavnice, koja u prvoj fazi ne bi bila reprezentativna za starohrvatski jezik, ali bi bila u potpunosti reprezentativna za tu sastavnicu i činila bi korpus za sebe. Tek s objavom novih sastavnica mogao bi se početi smatrati potkorpusom, a slika starohrvatskog jezika postajala bi jasnija i reprezentativnija.

Objavljivanje korpusa starohrvatskog jezika dalo bi toliko željene temelje Hrvatskoj jezičnoj riznici i nadamo se novi polet hrvatskoj filologiji u digitalnom dobu.

Literatura

Bowern, Claire. 2007. TshwaneLex Dictionary Compilation Software. *Language documentation & conservation* 1(1), 94—99.

Gau, Melanie. 2005. *The State of Historical Corpus Linguistics with Special Focus on the Russian Language.* M.A. thesis. University of Regensburg, Institute for Slavonic Languages and Literatures. *http://www.uni-r.de/Fakultaeten/phil_Fak_IV/Korpuslinguistik/meyer/PDF/melanie.pdf*, preuzeto 5.X.2010.

Joffe, David, Gilles-Maurice de Schryver, Daniel Jacobus Prinsloo. 2003. Computational features of the dictionary application »TshwaneLex«. *Southern African linguistics and applied language studies* 21(4), 239—250.

Kapetanović, Amir. 2007. Digitalizacija korpusa starohrvatskih tekstova i kritika teksta. U Sanja Seljan, Hrvoje Stančić, ur. *The Future of Informa-*

[36] »Within years, the DCNC should reach full representativeness for the period of the oldest Czech written records« (Kučera 1999:63).

tion Sciences: INFuture2007 — Digital Information and Heritage. Zagreb : Odsjek za informacijske znanosti, Filozofski fakultet. 173—182.

Kapetanović, Amir, Dragica Malić, Kristina Štrkalj Despot. 2010. *Hrvatsko srednjovjekovno pjesništvo : Pjesme, plačevi i prikazanja na starohrvatskom jeziku*. Zagreb : Institut za hrvatski jezik i jezikoslovlje. 944 str.

Kučera, Karel. 1999. The General Principles of the Diachronic Part of the Czech National Corpus. U V. Matoušek et al., ur. *Text, Speech and Dialogue*. Berlin, New York etc. : Springer. 62—65.

Kučera, Karel. 2007. Mapping the Time Continuum: A Major Raison d'être for Diachronic Corpora. U M. Davies et al., ur. *Proceedings of the Corpus Linguistics Conference CL2007*. University of Birmingham. http://ucrel.lancs.ac.uk/publications/CL2007/paper/27_Paper.pdf, preuzeto 5.X.2010.

Malić, Dragica. 2002. *Nacrt za Hrvatski rječnik do Marulića i njegovih suvremenika*. Zagreb : Institut za hrvatski jezik i jezikoslovlje. 154 str.

Manuscript Special Interest Group. *http://www.tei-c.org/SIG/Manuscripts/*, 1.X.2010.

McEnery, Tony, Andrew Wilson. 2001. *Corpus linguistics : An introduction*. Edinburgh : Edinburgh University Press. 224 str.

Meyer, Roland. 2005. The Regensburg Diachronic Corpus of Russian: A New Source for Linguistic Research (Not Only) on Modality. U B. Hansen, P. Karlík, ur. *Modality in Slavonic Languages: New Perspectives.* München : Sagner. 315—336.

Philologic. *http://sites.google.com/site/philologic3/home*, 1.X.2010.

Portada, Tomislav, Tomislav Stojanov. 2009. O vodoravnim crticama u hrvatskome pravopisu. *Filologija* 52, 91—120.

Schryver de, Gilles-Maurice, David Joffe. 2005. Dynamic Metalanguage Customisation with the Dictionary Application Tshwanelex. U F. Kiefer, G. Kiss, J. Pajzs, ur. *Computational Lexicography, COMPLEX 2005*. Budapest : Linguistics Institute, Hungarian Academy of Sciences. 190—199.

Schulte, Kim. 2009. Using non-annotated diachronic corpora: benefits, methods and limitations. U A. Enrique-Arias, ur., Diacronía de las lenguas iberorromances: nuevas aportaciones desde la lingüística de corpus, Madrid/Frankfurt : Iberoamericana/Vervuert. 167–182.

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Zagreb : Ex libris. 191. str.

TEI Lite. *http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html*., 1.X.2010.

TEI: P5 Guidelines. 2007. *http://www.tei-c.org/Guidelines*, 1.X.2010.

The Menota handbook. Version 2.0. *http://www.menota.org/guidelines/index.page*, 1.X.2010.

The TshwaneLex Suite. User Guide. Version 4.0.6. *http://tshwanedje.com/downloads/*, 1.X.2010.

Computer Designed Corpus of the Old-Croatian Language

Abstract

Research for the *Old-Croatian Dictionary* in the Department of Croatian Language History and Historical Lexicography at the *Institute of Croatian Language and Linguistics* consists of three main stages: textological, computational and lexicographic. The first stage (textological) is thoroughly explained in Kapetanović's article *Digitization of Old Croatian Texts and Textual Criticism*.

This paper explains the second stage — computational corpus design for the mentioned dictionary. Therefore, the analysis is focused on anticipating problems in designing a machine-readable corpus of texts written in the Old Croatian language, and on providing possible solutions. The computational processing of the texts for the corpus is described, as are problems which arise from the selection of *TEI* and *Philologic* server application. Priorities in corpus design are determined in order to overcome the mentioned problems and other potential difficulties. One of the possible solutions of separation of the *apparatus criticus* from the Old Croatian texts is proposed. All proposals are aimed for a more efficient encoding of primary data and metadata with respect to users, while remaining within the capabilities of the project.

The paper also deals with the third, lexicographic stage by giving an overview of the chosen lexicographic tool *TshwaneLex*.

Ključne riječi: starohrvatski jezik, korpus, standard TEI

Key words: Old-Croatian language, corpus, TEI standard